Exploratory Data Analysis

LIMITED PRELIMINARY EDITION

Volume III

by JOHN W. TUKEY

ADDISON-WESLEY PUBLISHING COMPANY
READING, MASSACHUSETTS
MENLO PARK, CALIFORNIA · LONDON · DON MILLS, ONTARIO

014.25

TABLE OF CONTENTS

Volume 3, Preliminary Edition

(For details, see tables of contents of individual chapters.)

Preface to volume 3

Introduction to the preliminary edition
About the problems

Guidelines for shortcutting and delaying

The structure of volume III

24. Expressing counts

25. Beginning to display counts (patterns of distribution)

26. Counts against a background

27. Counts of counts of instances

27x. Counts of counts of species [very long tails; rank-size-log technique]

28. Expressing counted fractions

29. One ordered response (mixed two-way analyses)

29x. Paired ordered responses -- counted (true contingency tables)

29y. Less simple examples

30. Unordered responses, many paired

30x. Ordering factors by relationship

30z. Diagonal Analysis

EXPLORATORY DATA ANALYSIS

VOLUME III

Limited Preliminary Edition

by

John W. Tukey

Princeton University
and
Bell Telephone Laboratories, Incorporated

The writing of this work, together with some of the research reported herein, was supported in part by the United States Government. Reproduction, translation, publication, use and disposal in whole or in part, by or for the United States Government is permitted.

Copyright 1970 by Addison-Wesley Publishing Company, Inc.
Philippines copyright 1970 by Addison-Wesley Publishing Company, Inc.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher. Printed in the United States of America. Published simultaneously in Canada.

TABLE OF CONTENTS

Volume 3, Preliminary Edition

(For details, see tables of contents of individual chapters.)

Preface to volume 3

Introduction to the preliminary edition
About the problems

Guidelines for shortcutting and delaying

The structure of volume III

24. Expressing counts

25. Beginning to display counts (patterns of distribution)

26. Counts against a background

27. Counts of counts of instances

27x. Counts of counts of species [very long tails; rank-size-log technique]

28. Expressing counted fractions

29. One ordered response (mixed two-way analyses)

29x. Paired ordered responses -- counted (true contingency tables)

29y. Less simple examples

30. Unordered responses, many paired

30x. Ordering factors by relationship

30z. Diagonal Analysis

PREFACE TO VOLUME 3

It is well to understand what you can do before you learn how to measure how well you seem to have done it. This is still true for counts and counted fractions. The other remarks in the Preface to Volume 1 apply with at least equal force to this volume -- please read them.

Techniques and concepts do build upon volumes 1 and 2, but even the extreme short-cut versions of those volumes (see after each page I-3) should provide a basis for going ahead.

It is a pleasure to acknowledge support, guidance, cooperation, and hard work. The Army Research Office (Durham), through a contract with Princeton University, and the Bell Telephone Laboratories have supported the writing financially. Charles P. Winsor taught the writer, during the 1940's, many things about data analysis that were not in the books. Although its formal beginning came after his death, it owes much to S. S. Wilks, whose leadership for statistics in Princeton made possible the gaining of the insights on which it is based. Careful reading, in various versions, by Dr. Leonard Sternberg and Mr. David Hoaglin have removed many errors, unclarities and awkwardnesses. (As always, the writer is solely responsible for those that remain.) Careful and skilled typing, mainly by Miss Elizabeth L. LaJeunesse, later to become Mrs. Elizabeth Dutka, and more recently by Mrs. Mary Bittrich, has been an essential.

It is important to read the Introduction to the Preliminary Edition which follows.

JOHN W. TUKEY

Princeton, New Jersey
January 17, 1971

INTRODUCTION TO THE PRELIMINARY EDITION

This is a slightly revised form of the third version of this account. Each version has been used in class in successive years. Each time enough has been learned to call for extensive revision and rearrangement. This was as true of the third trial as of the other two. This time, however, the signals are for a substantial reorientation rather than for the insertion of new material or the discussion of new techniques.

The preface has hinted at the separation of exploration from confirmation. What has not yet been recognized in this preliminary edition is the separation of breadth of possibility from seriousness of analysis. The First Edition, as seen dimly today, will try to bring the breadth of possibility into its first volume, making clear at the beginning some of the diversity of ways in which data can be explored. The second volume can then go back over these approaches in turn, explaining how each can be made more carefully -- made in a way suited to controversy or publication.

We do have three obligations to our data, obligations we ought to meet one after another. We need to explore the data flexibly -- here flexibility can be enhanced by simplicity without exhaustiveness. We need to carry out the chosen analyses carefully and exhaustively -- especially if we are to use them in a controversy. (All publication is potentially part of a controversy, just as all use as evidence is potentially part of an adversary position.) We need to assess their quality -- here is the place for Confirmatory Data Analysis.

The present version mixes exploration and care together, treating each major topic from both points of view before passing on to the next. The teacher may wish to move closer to the structure planned for the First Edition by going quite selectively through the three volumes of the preliminary edition. He cannot do this as well as he might; this version was not planned to let him do this -- but he can try. The guide to short-cutting that follows offers some help.

ABOUT THE PROBLEMS

The teacher needs to be careful about assigning problems. Not too many, please. They are likely to be longer than you think. The number supplied at the ends of the chapters is to accommodate diversity of interest, not to keep everyone busy.

Besides the length of our problems, both teacher and student need to realize that many problems do not have a single "right answer". There can be many ways to approach a body of data. Not all are equally good. For some this may be clear, but for other we may not be able to tell from this body of data which is preferred. Even several bodies of data about very similar situations may not be enough to show which should be preferred. Accordingly, it will often be quite reasonable for different analysts to reach somewhat different analyses.

Yet more -- to unlock the analysis of a body of data, to find the good way or ways to approach it, may require a key, whose finding is a creative act. Not everyone can be expected to create the key to any one situation. And, to continue to paraphrase Barnum, no one can be expected to create a key to each situation he or she meets.

An eminent statistician once visited an eminent prep school teacher of physics. After the laboratory group was well into the experiment, the statistician hinted that it might be interesting to collect all the students results and display them together. "Oh", said the teacher, "but some of them will have gotten the wrong answer." The usually free-spoken statistician, whom long experience had taught to expect variation in everything, dared not trust himself to reply. To learn about data analysis it is right that each of us try many things that do not work -- that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where we missed an opportunity -- at least until we were told about it. Each teacher needs to recognize this in grading and commenting on problems.

GUIDELINES FOR SHORTCUTTING AND DELAYING

Chapter	Various levels of shortcutting									
	Extreme		Intermediate		Moderate		Mild		None	
	Sections	*	Sections	*	Sections	*	Sections	*	Sections	*
24	A to D	11	A to D	11	A to D	11	A to G	17	A to G	17
25	A to D	13	A to D	13	A to D	13	A to D	13	A to E	18
26	A to D	12	A to E	13	A to E	13	A to G	17	A to G	17
27	A to B	8	A to B	8	A to B	8	A to F	16	A to F	16
27x	------	--	A to D	9	A to D	9	A to D	9	A to G	15
28	to 23	13	to 23	13	A to G	25	A to G	25	A to G	25
29	------	--	------	--	A to C	10	A to E	16	A to G	25
29x	A to D	16	A to D	16	A to E	22	A to F	28	A to G	29
29y	------	--	A to B	6	A to B	6	A to E	13	A to E	13
30	------	--	A to D	17	A to D	17	A to F	24	A to F	24
30x	------	--	------	--	A to D	10	A to D	10	A to D	10
30z	------	--	A to C	8	A to C	8	A to C	8	A to C	8
(total)		(73)		(114)		(152)		(184)		(215)

*Pages of text (not exhibits). Note that sections H and P, which would be covered, are omitted from the count.

THE STRUCTURE OF VOLUME III

This volume is not as clearly structured as one would like. Almost all of it has to do with counts, of one kind or another. The bulk of it has to do, not just with counts, but rather with counted fractions -- with counts thought of as parts of other larger counts. This distinction is important, and should guide all our attention to such topics.

When we can regard counts as telling us about one or more counted fractions, we are able to use a larger and more diversified kit of tools.

The only chapters that do not rest, in one way or another, dependent upon some of these tools, are chapters 24 and 30, which could, indeed, be absorbed into the earlier material.

Chapters 25 and 26, quite explicitly, and chapters 27 and 27x, less obviously, deal with pattern of distribution -- observed and fitted -- and their display. Counted fractions are in the background, rather than in the text, though the heavy use of ranks in chapter 27x comes extremely close to the explicit use of counted fractions.

The remaining chapters, 28, 29, 29x, 29y, 30x and 30z, at least make explicit use of counted fractions, and, when carefully examined, their analyses are seen not to be possible without some form of this notion.